Published online 1 April 2011

Nucleic Acids Research, 2011, Vol. 39, No. 11 ***e74***
doi:10.1093/nar/gkr130

ROS-DET: robust detector of switching mechanisms in gene expression

Mitsunori Kayano[1,2], Ichigaku Takigawa[1,2], Motoki Shiga[1,2], Koji Tsuda[2,3] and Hiroshi Mamitsuka[1,2,*]

[1]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji 611-0011, [2]Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Agency (JST) and [3]Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan

Received August 31, 2010; Revised February 5, 2011; Accepted February 22, 2011

ABSTRACT

A switching mechanism in gene expression, where two genes are positively correlated in one condition and negatively correlated in the other condition, is a key to elucidating complex biological systems. There already exist methods for detecting switching mechanisms from microarrays. However, current approaches have problems under three real cases: outliers, expression values with a very small range and a small number of examples. ROS-DET overcomes these three problems, keeping the computational complexity of current approaches. We demonstrated that ROS-DET outperformed existing methods, under that all these three situations are considered. Furthermore, for each of the top 10 pairs ranked by ROS-DET, we attempted to identify a pathway, i.e. consecutive biological phenomena, being related with the corresponding two genes by checking the biological literature. In 8 out of the 10 pairs, we found two parallel pathways, one of the two genes being in each of the two pathways and two pathways coming to (or starting with) the same gene. This indicates that two parallel pathways would be cooperatively used under one experimental condition, corresponding to the positive correlation, and the two pathways might be alternatively used under the other condition, corresponding to the negative correlation. ROS-DET is available from http://www.bic.kyoto-u.ac.jp/pathway/kayano/ros-det.htm.

INTRODUCTION

Gene expression analysis is a basic and important technique in molecular biology. There are two typical and simple concepts for expression analysis: (i) differential expression, which examines the difference in expression for a single gene between different experimental conditions (classes), such as case and control patients (1), and (ii) coexpression, which focuses on a combination of multiple genes, checking whether they are over- or underexpressed simultaneously (2). One notion with both of these two properties is differential co-expression, in which coexpression patterns differ depending upon the experimental conditions (3,4). We address an issue of finding one type of differential coexpression, which hereafter we call a 'switching mechanism'. The switching mechanism has two experimental conditions for expression of two genes, where two genes are positively correlated under one experimental condition while they are negatively correlated under the other condition (3,5–8). Figure 1 shows one simulated example of the switching mechanism. A simple, well-known case of the switching mechanism is Max, a transcription factor, which plays a role of an activator or a suppressor, depending on whether it binds to Myc (i.e. Myc-Max) or Mad (i.e. Mad-Max) (9). Another case is thyroid hormone receptor (TR), which forms a complex called TR-RXR and can be also an activator or a suppressor, depending on the absence or presence (amount) of thyroid hormone (10). Finding the switching mechanisms would be a key step to elucidating complex biological systems.

There are two typical techniques for finding switching mechanisms: the absolute difference of two correlation coefficients (3,5–7) and interaction test (8). Interaction test, being popular as a standard approach for detecting epistasis in genetics (11), is the log-likelihood ratio test between two logistic regressions with/without an interaction term. This test examines the strength of the interaction between two genes, which is in general equivalent to checking whether two genes can take the switching mechanism. However, a serious disadvantage of interaction test is its high computational burden, making it

*To whom correspondence should be addressed. Tel: +81 774 383023; Fax: +81 774 383037; Email: mami@kuicr.kyoto-u.ac.jp

hard to apply to a large number of gene combinations practically (8).

We then focus on the absolute difference of two correlation coefficients. The procedure of this approach is that we first compute the correlation coefficient of two genes in expression for each of the two experimental conditions and check the absolute difference of the two correlation coefficients. More concretely, in Figure 1, we first compute the correlation coefficient r_1 between gene X and gene Y for one experimental condition, say class 1 (shown by +), and r_2 between them for the other condition, say class 2 (shown by •). We then compute the absolute difference between these two correlation coefficients as the score of two genes X and Y as shown below:

$$s(X, Y) = |r_1 - r_2| \quad (1)$$

A switching mechanism must have this of a larger value, because two correlation coefficients should be different in the switching mechanism. We can consider any measure of correlation coefficients in this approach, such as the Pearson's correlation coefficient (5,7) and the Spearman's rank correlation (6) [(3) is a review over differential coexpression, including the switching mechanism]. Because of various correlation coefficients, there can be many approaches of using the absolute difference of two correlation coefficients, but they still have problems by which negative cases can be detected as positives. The negatives which can be detected incorrectly are caused by the following three main reasons: (i) outliers, (ii) range bias and (iii) a small number of examples. Figure 2a illustrates a typical negative case of (i), where only two points of class 2 are in the upper-right area, which are outliers and make this case pretend to be a switching mechanism. Figure 2b shows a case of (ii), where the range of expression values of class 2 is much smaller than that by class 1. Regardless of the very small range, class 2 can show a strong negative correlation in its figure, by which the absolute difference between the two correlation coefficients would become large. In reality, however, the small range of class 2 makes us unconvincing whether class 2 is negatively correlated or not. Thus, we cannot say that Figure 2b is a switching mechanism, and this means that Figure 2b should be a negative case. Figure 2c shows a typical case of (iii), where the number of points in two classes is so small that it cannot be considered as a switching mechanism, implying that this cannot be a positive gene pair.

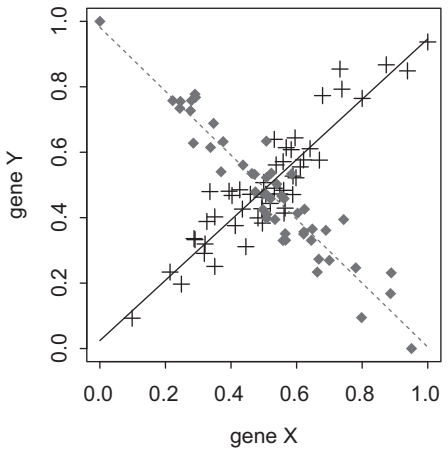

Figure 1. A sample of switching mechanisms with two classes (shown by plus symbol and filled square).

We propose a method, which we call ROS-DET (standing for RObust Switching mechanisms DETector) to avoid detecting the negative cases as positives (or to reduce the so-called false positive rates) in the paradigm of the absolute difference of two correlation coefficients. First, ROS-DET uses a robust measure of correlation coefficient, so-called biweight midcorrelation (12), to avoid detecting gene pairs with outliers such as Figure 2a as switching mechanisms. Most typical correlation coefficients robust against outliers are the Spearman's and Kendall rank correlations, but they use the ranking after sorting all the given values, which might lose some information in the original values. On the other hand, the biweight midcorrelation is a modification of the Pearson's correlation coefficient which allows to keep the original values and consider outliers carefully. The biweight midcorrelation was thus used in microarray data analysis already for detecting coexpression gene pairs (13), and the performance advantage of the biweight midcorrelation in expression analysis over other correlation coefficients was already shown (14). Second,

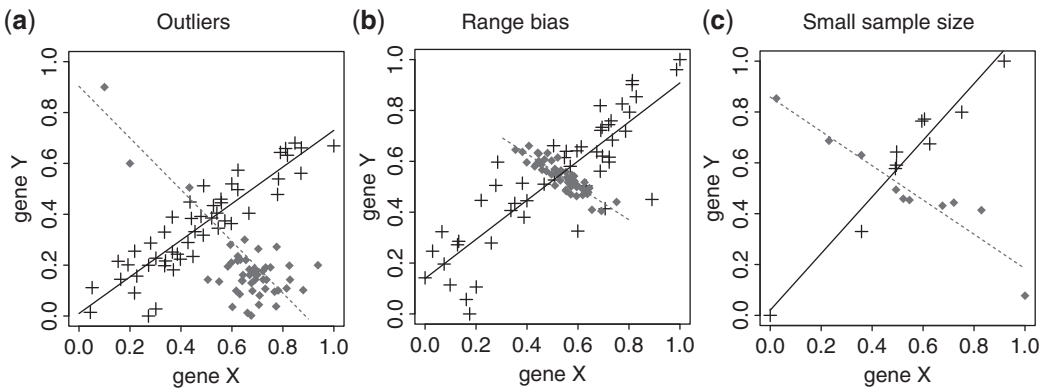

Figure 2. Three types of negative cases of switching mechanisms.

ROS-DET multiplies the difference of correlation coefficients by a weight, which reflects upon the range bias between two classes so that the weight becomes smaller as the range bias increases. This means that the absolute difference of correlation coefficients is lowered by the weight more as the range bias is larger. Third, ROS-DET uses *P*-values to discard gene pairs with only a small number of examples even if their weighted absolute difference of correlation coefficients is large. We use hypothesis testing to check the equality of biweight midcorrelation coefficients derived from two experimental conditions, meaning that gene pairs with higher *P*-values are more insignificant in difference of two biweight midcorrelation coefficients. On the whole, ROS-DET has two steps: WCOR (sorting gene pairs by the Weighted absolute difference between two biweight midCORrelation) and ECOR (the hypothesis testing on the Equality of two biweight midCORrelation). All gene pairs, which are generated from a given expression dataset, are first ranked by the weighted difference of two biweight midcorrelations in WCOR, and if *P*-values of gene pairs are high in ECOR, they are then deleted.

We experimentally confirmed the effectiveness of ROS-DET through experiments with synthetic and real datasets. We first generated synthetic datasets to examine the performance of WCOR on two problems, i.e. outliers and range bias, comparing to existing approaches. The results showed that WCOR outperformed all the other competing methods. We then checked the effectiveness of ECOR by using real data consisting of over 2.60×10^{10} gene combinations generated from 46 datasets in Gene Expression Omnibus (GEO). The result showed that removing gene pairs with high *P*-values relaxed the bias to the data with a smaller number of examples. Finally, we focused on the top 10 gene pairs in the final output of ROS-DET, to check the biological relevance of each pair. For 8 out of the 10 pairs, we found two (parallel) pathways, which start with or come to the same gene and, for each pair, have one of two genes at each pathway separately. The parallel pathways imply that two parallel pathways might be cooperatively used under one experimental condition, corresponding to the positive correlation of two genes, and the two pathways might be alternatively used under the other condition, corresponding to the negative correlation of two genes.

In summary, the main contribution of this article can be the following three: (i) we have developed ROS-DET based on the weighted midcorrelation coefficients, (ii) empirically validated the performance of ROS-DET and (iii) applied ROS-DET to real data to find new gene pairs with switching mechanisms.

MATERIALS AND METHODS

The main input of ROS-DET is an expression dataset measured under two experimental conditions, and the output is gene pairs with switching mechanisms between two experimental conditions. ROS-DET first generates all possible gene pairs from a given dataset and the following two steps are run over all pairs: (i) WCOR: all gene pairs are sorted by the weighted absolute difference of biweight midcorrelation coefficients and (ii) ECOR: out of the sorted gene pairs, pairs with high *P*-values regarding the equality of two biweight midcorrelation coefficients are removed. The feature of WCOR is biweight midcorrelation coefficient and the weighted absolute difference of two correlation coefficients, while that of ECOR is hypothesis testing on the equality of two biweight midcorrelation coefficients. Below, we explain each of the three features in detail and finally show the entire procedure of our approach.

Preliminaries

Let $\mathcal{D}$ be a given (microarray expression) dataset with G genes and two classes C_1 and C_2 (with N_1 and N_2 examples, respectively) corresponding to the two experimental conditions. We consider all possible pairs of genes over G genes. Let X and Y be two genes of an arbitrary pair. Let x_i be the expression value of gene X for an example i and similarly y_i be that of gene Y for i. Let $m_{x,k}$ be the median of expression values in class C_k for gene X, and $m_{y,k}$ be that for gene Y. Furthermore, let $\delta_{x,k}$ be the sample median of $|x_i - m_{x,k}|$ over all values x_i satisfying $i \in C_k$, and $\delta_{y,k}$ be that of $|y_i - m_{y,k}|$ over all y_i where $i \in C_k$. Let r_1 and r_2 be correlation coefficients for C_1 and C_2, respectively. When we already know the true value of the correlation coefficient, we write the true correlation coefficients by ρ_1 and ρ_2 for C_1 and C_2, respectively. For example, in order to generate synthetic data, we can use ρ_1 and ρ_2.

WCOR: biweight midcorrelation

The biweight midcorrelation r_k (12) for class C_k between two genes X and Y can be given as follows:

$$r_k = \frac{\sum_{i \in C_k} w(u_i) w(v_i)(x_i - m_{x,k})(y_i - m_{y,k})}{\sqrt{\left\{\sum_{i \in C_k} w^2(u_i)(x_i - m_{x,k})^2\right\}\left\{\sum_{i \in C_k} w^2(v_i)(y_i - m_{y,k})^2\right\}}},$$

where

$$w(z) = (1 - z^2)^2 \quad \text{if } |z| \leq 1$$
$$= 0 \qquad\qquad \text{otherwise,}$$

$u_i = (x_i - m_{x,k})/(K \cdot \delta_{x,k})$ and $v_i = (y_i - m_{y,k})/(K \cdot \delta_{y,k})$. [$K$ was set at nine in our experiments, according to (12,15,16).] Here $\sum_{i \in C_k} w^2(u_i)(x_i - m_{x,k})^2$ can be normalized into the 'biweight midvariance' (12), which is given in the Supplementary Data and hereafter we write by $q_{x,k}$. Similarly, we can write $q_{y,k}$ for the biweight midvariance of gene Y. We note that if weights $w(u_i)$ and $w(v_i)$ are always both 1 over all i and medians $m_{x,k}$ and $m_{y,k}$ are both the means (denoted by $\mu_{x,k}$ and $\mu_{y,k}$, respectively), then r_k is exactly the same as the Pearson's

correlation coefficient, which is given by:

$$r_k = \frac{\sum_{i \in C_k}(x_i - \mu_{x,k})(y_i - \mu_{y,k})}{\sqrt{\{\sum_{i \in C_k}(x_i - \mu_{x,k})^2\}\{\sum_{i \in C_k}(y_i - \mu_{y,k})^2\}}}.$$

WCOR: weighted absolute difference of two correlation coefficients

We modify Equation (1) into the following weighted form to deal with the range bias between two classes:

$$s(X, Y) = c|r_1 - r_2| \quad (2)$$

We here describe how we can compute c of Equation (2). The problem is to find a gene pair with a high bias (difference) in the range of expression values between two classes and then to discard it. To check the range of expression values, we can employ the variances, i.e. biweight midvariances, $q_{x,k}$ and $q_{y,k}$ for class C_k. We can say that the range of values is small if both $q_{x,k}$ and $q_{y,k}$ are small in class C_k. Thus, for class C_k we can keep the maximum of $q_{x,k}$ and $q_{y,k}$, and if this value is small, we can see that the range is small. The range bias can then be checked by the ratio of the range (the maximum variance) of one class to that of the other. In summary, we can first take the maximum of $q_{x,k}$ and $q_{y,k}$ for each class C_k and then take the ratio of the maximum variance of one class to that of the other class. Here, we note that the ratio can be in two ways, i.e. the ratio of class 1 to class 2 and the ratio of class 2 to class 1. We further note that Equation (2) must become lower as the range bias increases more, and so c in Equation (2) must take 1 for the case with no range bias and be reduced to zero as the range bias increases. Thus to make c hold this property, we can compute the possible two ratios and take the minimum of the two ratios.

Overall c can be given as follows:

$$c = \min\left\{\frac{\max\{q_{x,1}, q_{y,1}\}}{\max\{q_{x,2}, q_{y,2}\}}, \frac{\max\{q_{x,2}, q_{y,2}\}}{\max\{q_{x,1}, q_{y,1}\}}\right\} = \frac{\text{the larger variance in class 1 (or 2)}}{\text{the larger variance in class 2 (or 1)}}. \quad (3)$$

By using Equations (2) and (3), we can then compute the score of each gene pair, indicating the possibility that the gene expression of this pair can be a switching mechanism. Then by computing scores of all possible combinations of genes, WCOR can rank these combinations by the computed scores.

ECOR: hypothesis testing for the equality of two correlation coefficients

The purpose of the hypothesis testing is to check the equality of two correlation coefficients (each being derived from the corresponding one of two classes), meaning that the hypotheses are H_0: $\rho_1 = \rho_2$ and H_1: $\rho_1 \neq \rho_2$. Given the expression values of two genes with two classes, if we assume that they take a bivariate normal distribution for each class and the correlation coefficient is the Pearson's correlation coefficient for each class, test statistic T of the likelihood ratio test for H_0 and H_1 follows the chi-square distribution with one degree of freedom under H_0 (17,18):

$$T = \sum_{k=1}^{2} N_k \log\frac{(1 - r_k\hat{\rho})^2}{(1 - r_k^2)(1 - \hat{\rho}^2)} \sim \chi_1^2, \quad (4)$$

where $\hat{\rho}$ ($-1 < \hat{\rho} < 1$) is a maximum likelihood estimator for ρ_1 and ρ_2 under H_0, where $\rho_1 = \rho_2$, and $\hat{\rho}$ satisfies $\sum_{k=1}^{2} N_k(r_k - \hat{\rho})/(1 - \hat{\rho} \cdot r_k) = 0$. Interested readers should see the Supplementary Data for the derivation of Equation (4).

We apply this test statistic to the biweight midcorrelation. More concretely, we use the biweight midcorrelation given by Equation (2) for r_k in Equation (4). We show the empirical validation on this application of Equation (4) to the biweight midcorrelation in the Supplementary Data. We compute P-values by using this hypothesis testing with the biweight midcorrelation. We can say that two correlation coefficients of pairs with higher P-values than a specified significance level are not different statistically. Thus, out of the ranked list generated in WCOR, ECOR removes pairs with higher P-values than a pre-specified significance level.

The entire procedure

Given an expression dataset with two classes, ROS-DET first generates all possible pairs of G genes. WCOR then computes, for each of all pairs, the score given by Equation (2), and further sorts all pairs according to the computed scores. Finally, ECOR computes P-values of the sorted pairs by using Equation (4) to remove pairs with higher P-values than a prefixed significance level. Figure 3 shows a pseudocode of ROS-DET with WCOR and ECOR.

RESULTS

Validating WCOR (the first step of ROS-DET) with synthetic data

Experimental setting. The procedure of this experiment is that we first generate datasets according to some distribution, changing its parameters, by which some datasets are true positives and the rest are true negatives. That is, the problem setting is binary classification over the generated datasets, i.e. predicting whether each dataset is a positive or a negative. We then apply WCOR to this problem and compare its performance with those of the other competing methods.

For each dataset in class C_k, we generated examples according to a bivariate g-and-h distribution with four parameters, $g_k(\geq 0)$, $h_k(\geq 0)$, σ_k and ρ_k (19) [More concretely, we first generated examples according to bivariate normal distributions (for genes X and Y) and transformed them into those in polar coordinates, in which we used only the distance from the origin to each example by which we generated examples following an univariate g-and-h distribution. We show the detail of generating the bivariate g-and-h distribution in the Supplementary Data]. Here g_k controls the skewness of the distribution, and the distribution is symmetrical at

Require: $\mathcal{D}$: Expression dataset with G genes and two classes
N_{out}: Number of gene pairs to be output
α: Significance level for gene pairs to be selected
Ensure: $\mathcal{O}$: Gene pairs with their scores and P-values
ROS-DET($\mathcal{D}$, N_{out}, α)
// Preprocessing Step
$\mathcal{O} = \emptyset$;
for each gene X of G genes **do**
  for each class C_k $(k = 1,2)$ **do**
    Calculate midvariance $q_{x,k}$;
  end for
end for
// WCOR
for each gene X of G genes **do**
  for each gene Y of G genes **do**
    for each class C_k $(k = 1,2)$ **do**
      Calculate the biweight midcorrelation r_k according to Equation (2);
    end for
    Calculate weight c according to Equation (3);
    Calculate score $s(X,Y)$ according to Equation (2) and save the score;
  end for
end for
Sort all gene pairs according to their scores;
// ECOR
$n = 0$;
for each (X,Y) of the sorted gene pairs **do**
  Calculate the P-value, $p_{X,Y}$, by using test statistic T in Equation (4);
  if $p_{X,Y} < \alpha$ **then**
    $\mathcal{O} \leftarrow \mathcal{O} \cup ((X,Y), p_{X,Y}, s(X,Y))$;
    $n \leftarrow n+1$;
  end if
  if $n = N_{\text{out}}$ **then**
    Break;
  end if
end for

Figure 3. Pseudocode of ROS-DET for detecting gene pairs which are most likely to have switching mechanisms.

$g_k = 0$, which was used throughout all our experiments, meaning that this parameter had nothing to do with our experiment. h_k, a non-negative parameter, controls the kurtosis (peakedness) of the distribution. We note that if $g_k = h_k = 0$, the distribution is the standard normal distribution. As h_k becomes larger, the distribution becomes broader (or the distribution tail becomes more emphasized), meaning that outliers can be generated more easily. σ_k and ρ_k are the variance of the distribution and the (true) correlation coefficient between two variates (genes X and Y), respectively. We note that σ_k can be used to generate the dataset with a different range between genes X and Y. Under a fixed h_k, we change σ_k and/or ρ_k according to some manner, where under each parameter setting, we generate a certain number of datasets.

We tested three values of h_k: 0, 0.5 and 1, fixing $N_1 = N_2 = 100$. Under each of them, we changed σ_k and/or ρ_k in the following three types of manners:

(1) Randomness: under a certain h_k, we generated datasets, changing ρ_k such that $\rho_1 = -\rho_2 = 0, 0.01, \ldots, 1$, fixing $\sigma_1 = \sigma_2 = 1$. $\rho_k = 0$ corresponds to that of no correlations among examples in C_k, while $\rho_k = 1$ corresponds to the state that examples are totally correlated between two variates in C_k. This means that examples are totally random when $\rho_1 = -\rho_2 = 0$ and examples must be the switching mechanism when $\rho_1 = -\rho_2 = 1$. Thus, datasets with ρ_1 and ρ_2 closer to $\rho_1 = -\rho_2 = 1$ should be positives and the rest should be negatives. In fact, we used ρ_k for r_k in Equation (5), and $\hat{\rho}$ in Equation (5) can be computed from each generated dataset, and then if the P-value was less than the significance level (5%), we assigned 'positive' to the true class label of the dataset; otherwise, 'negative' was assigned.
In prediction, we again emphasize that we used WCOR only, in which we computed score $s(X, Y)$ in Equation (2) for each dataset and sorted all datasets by scores.
We generated datasets 100 times for each pair of ρ_1 and ρ_2 and the results were averaged over total runs under each setting.

(2) Range bias: under a certain h_k, we generated datasets, changing the ratio of ranges (variances) between two classes such that $\sigma_2 = 1, 1.1, \ldots, 10$, fixing $\sigma_1 = 1$ and $\rho_1 = -\rho_2 = 0.5$. Thus, $\sigma_2 = 1$ shows that the ratio is 1, while $\sigma_2 = 10$ shows that the ratio is 10 (or $\frac{1}{10}$). In this case, datasets with lower σ_2 should be positives and those with higher σ_2 should be negatives. The true class label is assigned and predicted in a similar manner in the previous experiment, but in true class label assignment, we used the Bartlett test (20), which is the hypothesis testing on the equality of four different variances, to be generated in our case by the combinations of two classes and genes X and Y. We used the significance level at 5% for the P-value of the Bartlett test, meaning that a dataset was positive if its P-value is higher than 5%. We generated datasets 100 times for each σ_2 and the results were averaged over total runs under each setting.

(3) Randomness + Range bias: we changed both ρ_k and σ_k here so that $\rho_1 = -\rho_2 = 0, 0.01, \ldots, 1$, $\sigma_1^2 = 1$ and $\sigma_2^2 = 1, 1.1, \ldots, 10$. To assign a class label to each dataset, we used both Equation (4) and the Bartlett test and regarded the datasets which were labeled as positives by both of these two tests as positives and the rest as negatives. We generated datasets 10 times for each parameter setting and the results were averaged over total runs under each h_k.

For each of the above three settings, we show the distribution of data points of representative samples in the Supplementary Data.

Results. We evaluated the performance of each competing method by using receiver operator characteristics (ROC) curves. Figure 4 shows the average ROC curves of WCOR (colored by red), being compared with those of four competing methods: (i) the absolute difference of two correlation coefficients when using the Pearson's correlation coefficient (Pearson), (ii) the Spearman's correlation coefficient (Spearman), (iii) the (unweighted) biweight midcorrelation (Biweight) and (iv) the interaction

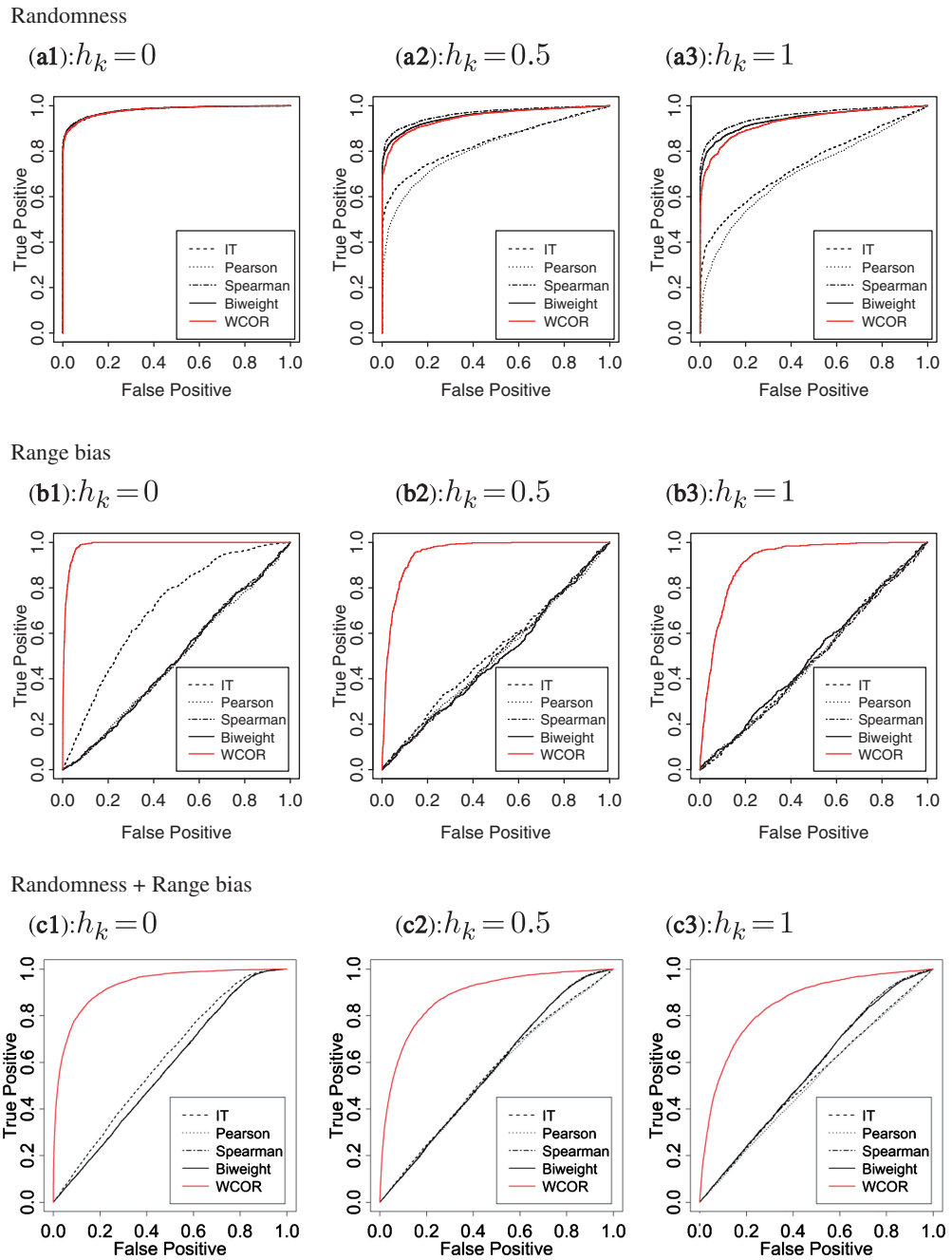

Figure 4. ROC curves by WCOR (colored red), comparing those by interaction test (IT), the absolute difference between two Pearson's correlation coefficients (Pearson), that between two Spearman's correlation coefficients (Spearman) and that between biweight midcorrelation coefficients (Biweight).

test (IT). Figure 4a1–a3 show the results of Randomness for $h_k = 0$, 0.5 and 1.0, respectively, where outliers were increased as h_k increased. Figure 4a1 shows that all ROC curves were lowered as increasing h_k as shown by Figure 4a2 and a3 except those using robust correlation coefficients, i.e. WCOR, Spearman and Biweight, implying that these methods were robust against outliers. Figure 4b1–b3 show the results of Range bias for $h_k = 0$, 0.5 and 1.0, respectively. These results show that WCOR outperformed other competing methods clearly, for any amount of outliers. In particular, the performance advantage of WCOR over other methods was clear in Figure 4b2 and b3, where the performances of the other methods were almost on the diagonal line, implying that their performances were similar to random guessing while WCOR showed good performance by showing a typical ROC curve. Figure 4c1–c3 show the results of Randomness + Range bias, which is a more real situation, for $h_k = 0$, 0.5 and 1.0, respectively. This case, again, WCOR outperformed other four competing methods clearly for all cases as shown by Range bias. Overall, WCOR outperformed other methods in a more

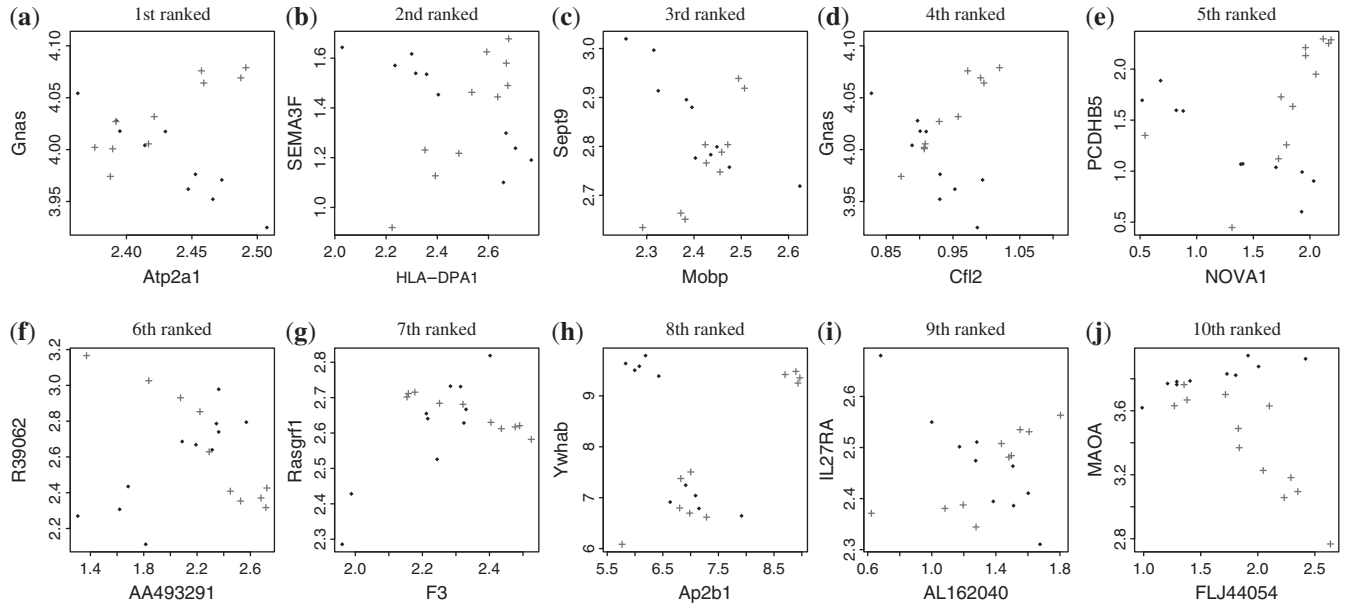

Figure 5. Expression values of the top 10 gene pairs in the output of WCOR for real data.

real situation, keeping its performance under Randomness as a comparable one to noise-robust methods, such as the Spearman's rank correlation and the biweight midcorrelation. We run our experiments over the cases with $N_1 = N_2 = 50$ and $N_1 = N_2 = 20$ and confirmed that the performance advantage of WCOR over other competing methods was all kept, and this result is attached to the Supplementary Data. Furthermore, we checked the case with $N_1 = N_2 = 10$ (shown in the Supplementary Data), where however the advantage of WCOR was not significant but this case was mostly removed by ECOR in real situations, implying that this result will not affect the reliability of ROS-DET.

Validating ECOR (the second step of ROS-DET) with real data

Experimental setting. Out of 2089 GEO DataSets (GDSs) of the GEO database (21) of the latest update in July, 2008, we extracted 46 datasets (or GDSs) which satisfy the following two conditions:

(1) Experimental conditions can be divided into two or more classes.
(2) Each class has 10 or more experiments.

The 46 datasets are listed in the Supplementary Data. Using the 46 datasets, ROS-DET first generated 2.60×10^{10} gene combinations. WCOR sorted out 2.60×10^{10} pairs, according to the weighted absolute difference of biweight midcorrelation. ECOR then discarded gene pairs with *P*-values higher than a prefixed threshold which was set to $1.92 \times 10^{-12} \simeq 0.05/(2.60 \times 10^{10})$ by considering the Bonferroni correction and the significance level of 5%. We used the Bonferroni correction, which is known to be relatively conservative, to keep reliable gene pairs only. The top 100 gene pairs by ECOR, i.e. the final output, is shown in the Supplementary Data.

We first compared the distribution of expression values of the top 10 gene pairs by WCOR with that by ECOR. We then focused on gene pairs, each being derived from a GDS with, for each class, a certain number of replicates (experiments) which we call 'the number of examples' or 'sample size'. We showed the distribution (histogram) of the number of examples, being generated by the top 1000 gene pairs by WCOR, comparing with those by the original dataset and the top 1000 gene pairs by ECOR.

Results. Figure 5 shows expression values under two experimental conditions (• and +) of the top 10 gene pairs in the output of WCOR. This figure reveals that the number of examples was very small for any gene pair, by which some pairs cannot necessarily be switching mechanisms. For example, the eighth ranked pair consisted of three distant islands, which could not have been a switching mechanisms, and the 10th ranked pair had two non-overlapped distributions which also could not have been a switching mechanism. On the other hand, Figure 6 shows the top 10 gene pairs in the output of ECOR. Each distribution of Figure 6 can be seen as a switching mechanism more clearly than those of Figure 5. This result indicates that the outputs of WCOR were likely to be the cases with the smaller number of examples, and ECOR works for removing dubious cases in the outputs of WCOR.

Figure 7a shows the distribution of the number of examples (sample sizes), by the original 46 datasets. This distribution is rather uniform, meaning that 46 datasets have a variety of sample sizes relatively equally, in the range of 10–70. Figure 7b shows the distribution by the top 1000 gene pairs obtained from WCOR. We here note that the total sum of the number of all gene pairs in (Figure 7b) is 2000 ($= 1000 \times 2$), since each gene pair

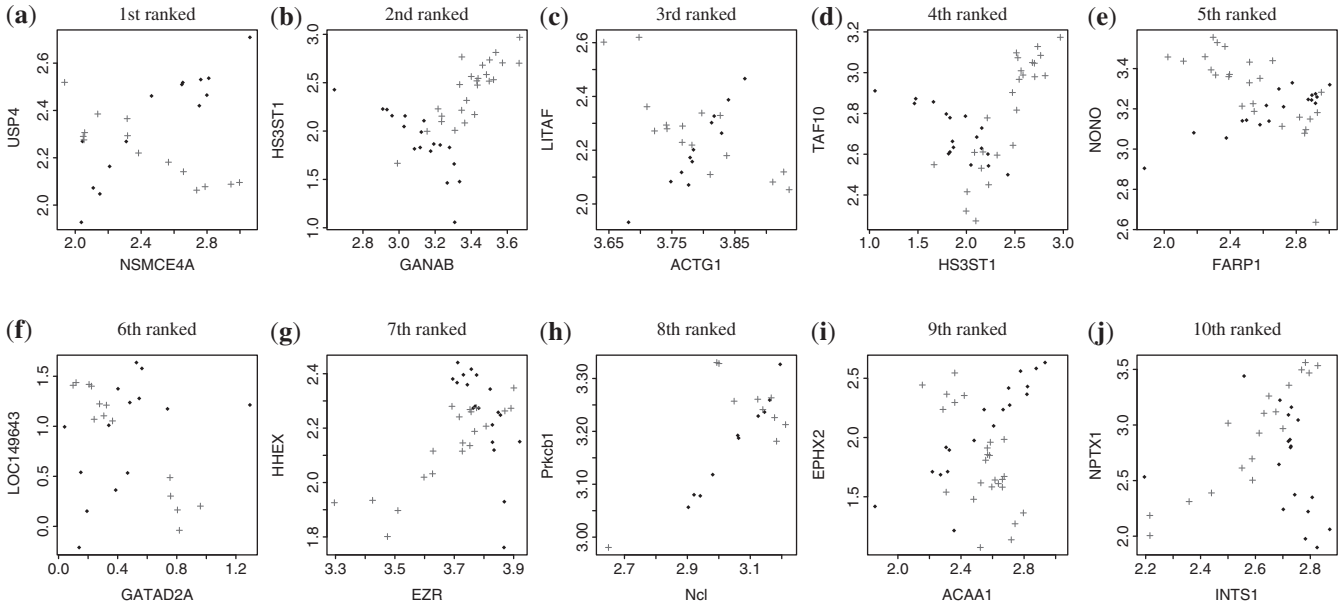

Figure 6. Expression values of the top 10 gene pairs in the output of ECOR, i.e. the final output of ROS-DET, for real data.

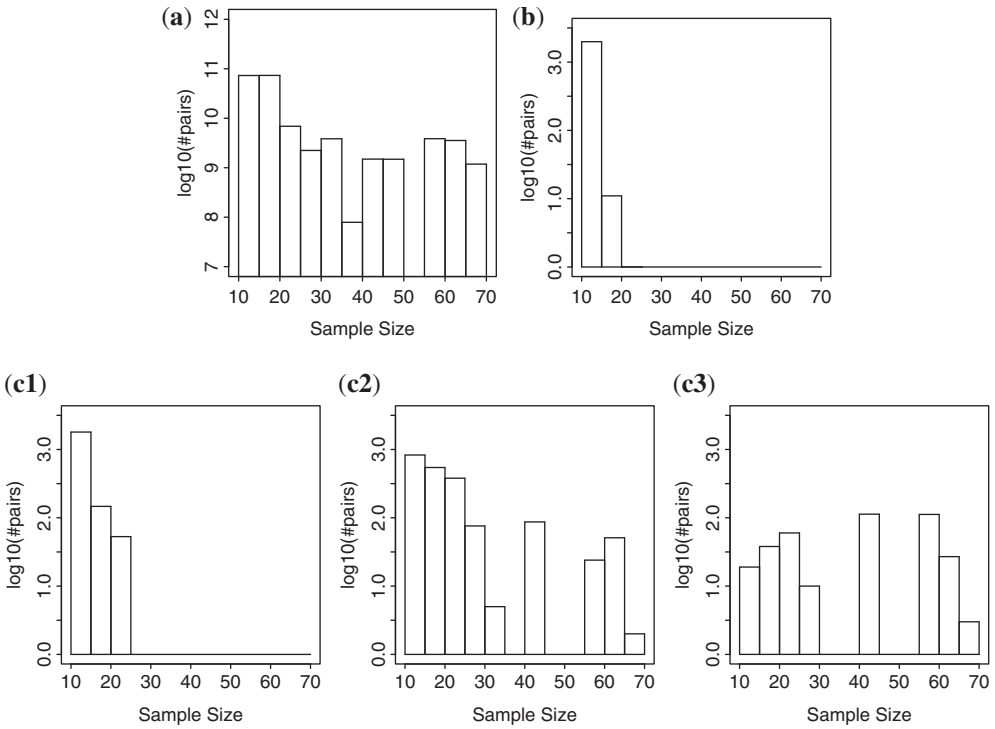

Figure 7. The log of #pairs, denoted by log10(#pairs) versus the number of examples, denoted by sample size, for (**a**) original data, (**b**) the top 1000 in the output of WCOR and (**c**) the top 1000 in the output of ECOR (ROS-DET), where the significance level was set at (**c**1) $1.0e-8$, (**c**2) $1.0e-9$ and (**c**3) $1.92e-12$, which was actually used in ROS-DET.

has two classes and the top 1000 gene pairs are obtained for each class. This figure clearly reveals that the top 1000 gene pairs by WCOR were so biased, where the number of examples was all less than 20, being consistent with the figures in Figure 5, all with only a small number of examples. We then show that this bias can be relaxed by ECOR. Figure 7c shows the distribution by the top 1000 gene pairs by ECOR, when we changed the cut-off value for *P*-values from 10^{-8} (Figure 7c1) to 10^{-9} (Figure 7c2) and further to 1.92×10^{-12} (Figure 7c3) [Again note that that the total sum of *y*-axis of Figure 7c1 and Figure 7c2 is equal to 2000 $(=1000 \times 2)$. Further, note that the

Table 1. Detail of the top 10 gene pairs outputted from ROS-DET by using real data

	Score	*P*-value	Gene pair	GDS	No. of examples in classes 1 and 2	Annotation (Two classes: •/+)
1	1.710	1.56×10^{-12}	{NSMCE4A,USP4}	GDS2656	14,14	Fetal/adult
2	1.688	2.77×10^{-13}	{HS3ST1,GANAB}	GDS2545	18,25	Normal prostate tissue/ metastatic prostate tumor
3	1.684	1.30×10^{-12}	{ACTG1,LITAF}	GDS1726	12,16	Control/HIV encephalopathy
4	1.639	8.29×10^{-13}	{HS3ST1,TAF10}	GDS2545	18,25	Normal prostate tissue/ metastatic prostate tumor
5	1.606	1.31×10^{-12}	{FARP1,NONO}	GDS2545	18,25	Normal prostate tissue/ metastatic prostate tumor
6	1.584	9.73×10^{-13}	{GATAD2A,LOC149643}	GDS1917	14,14	Control/schizophrenia
7	1.581	2.68×10^{-13}	{EZR,HHEX}	GDS1650	19,20	Adjacent normal/tumor
8	1.532	8.05×10^{-13}	{Ncl,Prkcb1}	GDS1455	10,10	Medial motoneuron/ intermediolateral column motoneuron
9	1.522	3.66×10^{-13}	{ACAA1,EPHX2}	GDS2545	18,25	Normal prostate tissue/ metastatic prostate tumor
10	1.488	1.64×10^{-12}	{INTS1,NPTX1}	GDS963	18,18	Normal/macular degeneration

distribution of Figure 7c3 is shown by 191 gene pairs only, which were all pairs obtained at this cut-off.]. The distribution in Figure 7c1 was still biased and very similar to Figure 7b, while that of Figure 7c2 was rather similar to Figure 7a, meaning that the bias in Figure 7c1 was relaxed by lowering the cut-off value. Finally, Figure 7c3 also shows an uniform distribution, though the shape is awkward because of the small number of gene pairs. We note that Figure 7c3 was obtained by a systematic manner in which we used the significance level of 0.05 as a cut-off value after applying the Bonferroni correction to *P*-values. Overall, these results confirmed that ECOR relaxed the bias in the result of WCOR by removing dubious pairs.

Validating highly ranked gene pairs with the literature

Table 1 shows the information of the top 10 gene pairs ranked by ROS-DET. For example, • and + in Figure 6 correspond to the left- and right-hand sides, respectively, of annotations in Table 1. In Table 1, two genes of each pair, say genes *X* and *Y*, have a clear switching mechanism in gene expression, to be related with some biological reason. Thus, we could make a pathway of genes that connects genes *X* and *Y*, where each step of the pathway would show a biological function like 'binding' or 'positive regulation'. This case, one naturally arising question is what biological system causes a switching mechanism, i.e. that genes *X* and *Y* are correlated under one condition and negatively correlated under the other condition. This can be explained by the following mechanism, which we call 'parallel pathways'.

Parallel pathways. Parallel pathways are two pathways which both reach (or start with) the same gene, which we call the destination gene, and these pathways satisfy the following two conditions: (i) gene *X* is in one pathway and gene *Y* is in the other, and (ii) under one experimental condition, genes *X* and *Y* are positively correlated in expression (meaning that two pathways are cooperatively used), and under the other experimental condition, genes *X* and *Y* are negatively correlated in expression (meaning that two pathways are alternatively used). One possible scenario of a switching mechanism with parallel pathways is that the expression of the destination gene can be controlled by (or can control) the two conditions. That is, the expression of the destination gene can be changed by (or can change) the cooperative or alternative expression of upstream (or downstream) genes.

One typical example found by Li (5) has two genes, each being in a different metabolic pathway from glutamate to ornithine, controlled by the expression of CPA2, which is in upstream of glutamate. That is, if the expression of CPA2 is low, two genes are positively correlated in expression, while if that of CPA2 is high, two genes are negatively correlated, meaning that these two genes are alternatively expressed.

Figure 8 illustrates a simulated example of two genes, genes X and Y, with a switching mechanism and parallel pathways, where gene Z is the destination gene. We further note that genes X and Y must be correlated with the gene Z in a particular manner: genes X and Y should be both positively or negatively correlated with gene Z under one condition (corresponding to + in Figure 8 showing both positive correlations), while genes X and Y should be correlated with gene Z in two different ways, i.e. one being positive and the other being negative, under the other condition (corresponding to • in Figure 8).

We here check how frequently the parallel pathways can be found in the top 10 gene pairs ranked by ROS-DET and further confirm their validity by computing the biweight midcorrelations between highly ranked paired genes and the destination genes.

Top ranked gene pair. NSMCE4A [or NSE4A: non-SMC element 4 homolog A (*Saccharomyces cerevisiae*)] and USP4 (ubiquitin specific peptidase 4 or UNP) from GDS2656.

NSMCE4A has a human homolog, NSE4B (or EID3). The EID family has EID1, EID2 and EID3, in which both EID1 and EID2 inhibit cell differentiation while both

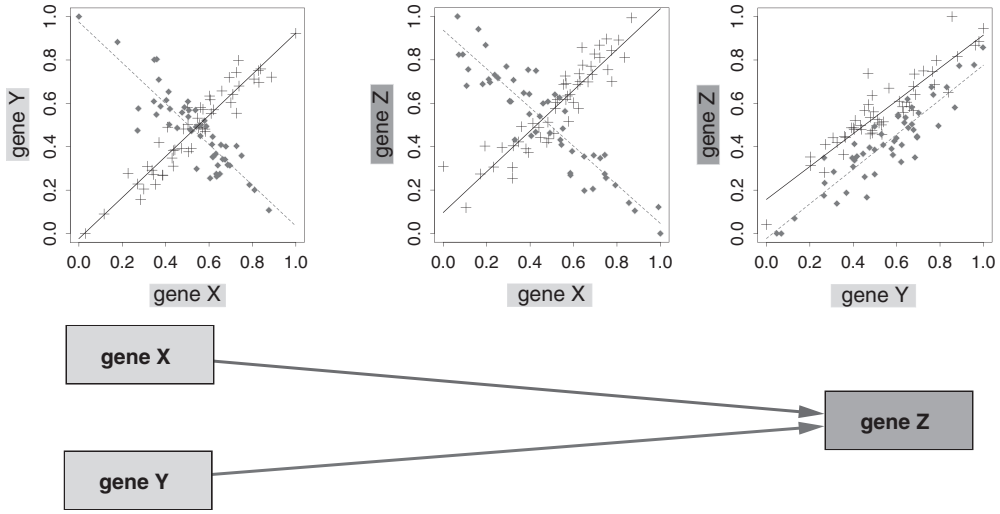

Figure 8. Switching mechanism with parallel pathways.

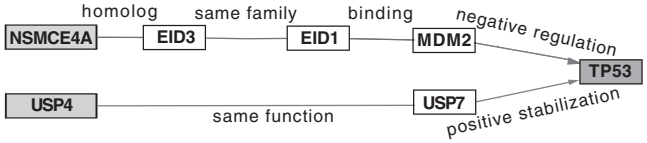

Figure 9. Pathways for the top ranked gene pair.

EID1 and EID3 inhibit transcription under the existence of EP300 (22). This makes us assume that NSMCE4A has the same biological function as EID1. EID1 interacts with (probably binds to) MDM2 (23), which is a negative regulator of tumor suppressor protein TP53 (24), implying that MDM2 is rather a positive factor on tumor growth. On the other hand, USP7, a ubiquitin specific peptidase with the same function as USP4, stabilizes the activation of TP53 (25), implying some negative effect on tumor growth.

Figure 9 summarizes these genes into two parallel pathways which both go to TP53. This figure shows a direct implication to the mechanism between EID1 and USP4 (or USP7), being consistent with parallel pathways. Two conditions of GDS2656 are fetal and adult. Figure 6a shows that two genes are positively correlated under fetal while negatively correlated under adult, implying that TP53 might be alternatively regulated by NSMCE4A and USP4 under adult, while TP53 could be regulated in a cooperative manner under fetal. To confirm this finding, we attempted to check the expression of TP53 but we could not find TP53 in GDS2656, and so instead we checked the expression of TP53RK, a TP53 regulating kinase. Table 2 shows the biweight midcorrelation between TP53RK and each of two genes of the top ranked pair. The result shows that TP53RK is positively correlated with both NSMCE4A and USP4 under fetal, while under adult, TP53RK is positively and negatively correlated with NSMCE4A and USP4, respectively. This result is consistent with our finding, which implies cooperative regulation under fetal and alternative regulation under adult.

Table 2. Biweight midcorrelations in gene expression of the top ranked pair

Gene 1	Gene 2	r_1(•: Fetal)	r_2 (+: Adult)
NSMCE4A	TP53RK	0.25	0.50
USP4	TP53RK	0.25	−0.48
MDM2	TP53RK	0.35	0.40
USP7	TP53RK	−0.27	−0.40

We further checked the biweight midcorrelation in expression between two neighboring genes which are in the parallel pathways and indicated by positive or negative regulation, e.g. negative regulation of MDM2 on TP53. We here note that we did not check neighboring pairs labeled by 'binding', since binding is elusive by including both positive and negative regulation. Thus, the neighboring pairs we checked are MDM2 → TP53 (negative regulation) and USP7 → TP53 (positive stabilization), and we used TP53RK instead of TP53. Table 2 shows the result, indicating that under both fetal and adult TP53RK was positively and negatively correlated with MDM2 and USP7, respectively, which are both reverse to the labels assigned to MDM2 → TP53 and USP7 → TP53. This implies that TP53RK might be negatively regulated by TP53, and this finding is consistent with (26). We note that even if TP53RK is negatively regulated by TP53, it is not contradictory to our finding that NSMCE4A and USP4 are cooperatively correlated under fetal and are alternatively correlated under adult.

Second ranked gene pair. HS3ST1 [heparan sulfate (glucosamine) 3-O-sulfotransferase 1] and GANAB (α-glucosidase) from GDS2545.

Both of these two genes are related with generating glycans (27,28). HS3ST1 is involved with cancer cells (28–30), while GANAB is related with carbohydrate absorption and digestion, especially metastatic process

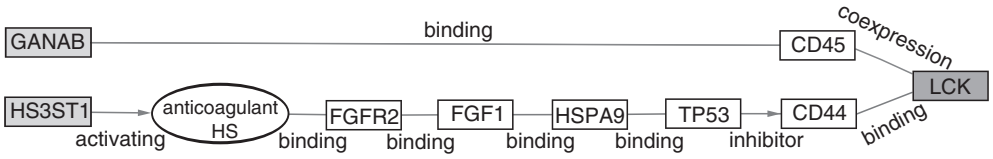

Figure 10. Pathways for the second ranked gene pair.

of tumor cells (31). Two GANAB inhibitors, 1,6-epi-cyclophellitol and castanospermine, inhibit experimental metastasis of tumor and tumor growth (31,32), meaning that GANAB may be rather an activator for tumor (metastasis), since these two GANAB inhibitors prevent tumor growth. GANAB stably interacts with (probably binds to) the external domain of PTPRC (CD45) (33), where Lck is completely dysfunctional in the absence of CD45 (34), and CD44 binds to Lck (35). On the other hand, HS3ST1 activates anticoagulant heparan sulfate (HS) (36) and can elevate the level of anticoagulant HS (37). Anticoagulant HS binds to FGFR2 (36), which further binds to FGF1 to produce a protein complex (38). FGF1 binds to HSPA9 (39), which binds to TP53 (40), an inhibitor of CD44 (41). CD44 selectively associates with active Lck, meaning that CD44 and Lck can form a complex.

Figure 10 summarizes these relations into parallel pathways both reach to the complex of CD44 and Lck. Two classes of GDS2545 are normal tissues and prostate tumor tissues. Figure 6b shows that GANAB and HS3ST1 are negatively correlated in expression under normal tissues, while they are positively correlated under tumor tissues. Here, TP53 is the inhibitor of CD44 in the pathway, by which some switching mechanism might be generated between GANAB and HS3ST1. That is, under normal tissues, the negative correlation between GANAB and HS3ST1 might lead to Lck and CD44 being highly expressed when the expression of one of HS3ST1 and GANAB is high (and the other is low). On the other hand, under tumor tissues, the correlation in expression between GANAB and HS3ST1 might indicate that the expression of CD44 and Lck is not well balanced. For example, even if both GANAB and HS3ST1 are well expressed, one of CD44 and Lck might be poorly expressed despite of the high expression of the other, which might be a result of the disorder, i.e. prostate tumor. To confirm this inference, we checked the biweight midcorrelation in expression between each gene of the second ranked pair and Lck (and CD44). Table 3 shows that under normal tissues Lck (and CD44) is correlated negatively and positively with GANAB and HS3ST1, respectively, implying that both CD44 and Lck can be expressed well when HST3ST1 is expressed highly and GANAB is expressed poorly. On the other hand, under tumor tissues, the expression of CD44 and Lck is unbalanced under any situation in expression of GANAB and HS3ST1. This result is clearly consistent with the above inference.

We further checked the biweight midcorrelation of neighboring two genes, which are in the parallel pathways and expected to be positively or negatively

Table 3. Biweight midcorrelations in gene expression of the second ranked pair

Gene 1	Gene 2	r_1(•: Normal)	r_2 (+: Tumor)
GANAB	Lck	−0.38	−0.44
HS3ST1	CD44	0.28	0.34
HS3ST1	Lck	0.24	−0.47
CD45	Lck	−0.12	0.55
TP53	CD44	0.50	−0.40

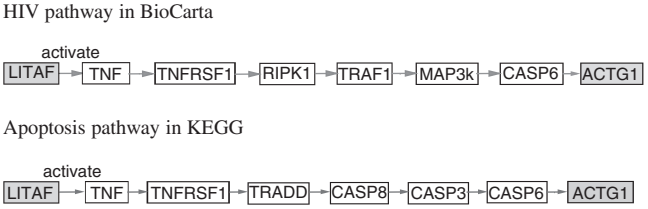

Figure 11. Pathways for the third ranked pair.

coexpressed. Again we did not check pairs labeled by 'binding', since binding includes both positive and negative regulation. We then checked CD45 → Lck (coexpression) and TP53 → CD44 (inhibitor). Table 3 shows that under tumor tissues, CD45 and Lck are positively correlated and TP53 and CD44 are negatively correlated, being consistent with the labels assigned to these two edges in Figure 10.

Third ranked gene pair use. ACTG1 (actin gamma 1) and LITAF [lipopolysaccharide-induced tumor necrorosis factor (TNF)-α factor] from GDS1726.

LITAF activates the production of infection-fighting substance called TNF-α. Both ACTG1 and TNF can be found in a HIV pathway (42), where TNF is an upstream signal while ACTG1 appears in downstream. In fact, the two experimental conditions measured in GDS1726 are the HIV encelophathy and its control. This pathway appears in the apoptosis pathway of KEGG. Figure 11 shows the two corresponding pathways. In both of these pathways, simply LITAF or TNF is in upstream and ACTG1 is in downstream, implying that this pair is not parallel pathways, but TNF and ACTG1 were connected by two different pathways, implying that this pair might be explained by another parallel association, which might cause a switching mechanism.

Fourth ranked gene pair. HS3ST1 [heparan sulfate (glucosamine) 3-*O*-sulfotransferase 1] and TAF10 [RNA polymerase II and TATA box binding protein (TBP)-associated factor] from GDS2545.

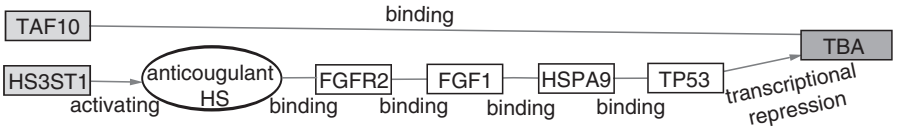

Figure 12. Pathways for the fourth ranked gene pair.

HS3ST1 is one of two genes appeared in the second ranked gene pair. We can then use the same pathway as that of the second ranked gene pair, and TP53 binds to a TBP and represses its transcription (43). On the other hand, TAF10 forms a complex (TFIID) with TBP and other proteins, where TAF10 is important for stabilizing the complex. We can summarize these genes into Figure 12, which shows that HS3ST1 and TAF10 have parallel pathways. Two conditions of GDS2545 are normal tissues and prostate tumor tissues. Figure 6d shows that TAF10 and HS3ST1 are negatively correlated in expression under normal tissues, while they are positively correlated under tumor tissues. Here, TP53 represses transcription of TBP, with which TAF10 forms a complex, by which under normal tissues, the negative correlation between TAF10 and HS3ST1 might indicate that both TBP and TAF10 are highly expressed when the expression of HS3ST1 is low (and that of TAF10 is high), while they are both not expressed highly when the expression of HS3ST1 is high (and that of TAF10 is low). On the other hand, under tumor tissues, the expression of TAF10 and that of HS3ST1 are positively correlated, possibly implying that the balance in expression between TBP and TAF10 (by which a complex will be formed) is not kept well, maybe because of the disorder, i.e. prostate tumor. In order to confirm this inference, we checked the biweight midcorrelation between TBP and two genes in the fourth ranked pair, and Table 4 shows the result. From this table, we can see that under normal tissues, TBP is negatively and positively correlated with HS3ST1 and TAF10, respectively, being consistent with our scenario. That is, under normal tissues, both TBP and TAF10 can be expressed highly when HS3ST1 is not expressed well, while TBP and TAF10 will not be expressed highly if HS3ST1 is expressed well. On the other hand, under tumor tissues, TBP can be positively correlated with both HS3ST1 and TAF10, although the correlation values are relatively slight, indicating that the above scenario or mechanism under normal tissues would not work well under tumor tissues.

We further checked the biweight midcorrelation between two neighboring genes, being labeled by positive or negative regulation (as mentioned earlier, we did not check those labeled by 'binding'). We then checked TP53 $\rightarrow$ TBP (transcription repression) (since anticoagulant HS is a chemical compound and not in GDS2545). Table 4 shows the biweight midcorrelation between TP53 and TBP, indicating negative correlation under both normal and tumor tissues, which is consistent with the fact that TP53 represses TBP.

Table 4. Biweight midcorrelations in gene expression of the fourth ranked pair

Gene 1	Gene 2	r_1(•: Normal)	r_2 (+: Tumor)
HS3ST1	TBP	−0.46	0.18
TAF10	TBP	0.43	0.26
TP53	TBP	−0.59	−0.33

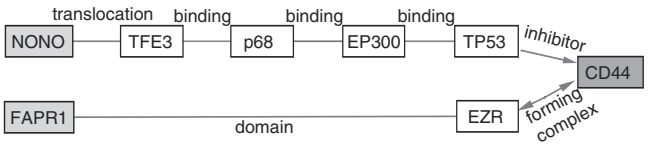

Figure 13. Pathways for the fifth ranked gene pair.

Fifth ranked gene pair. FARP1 [FERM RhoGEF (ARHGEF) and pleckstrin domain protein] and NONO (non-POU domain containing octamer-binding) from GDS2545.

NONO is deeply involved with human diseases, being a cause of cell carcinoma by a translocation with TFE3. TFE3 binds to a tumor suppressor p68 (DDX5) (44), which binds to EP300 (45), which further binds to TP53 (46). TP53 is an inhibitor of CD44 (41). On the other hand, CD44 forms a complex with ezrin family proteins (47). FARP1 has three domains including an ezrin-like domain. Figure 13 summarizes these genes into pathways which have CD44 as their final destination, indicating parallel pathways. Two classes are normal tissues and prostate tumor tissues, which are the same as those of the second and fourth ranked gene pairs. Figure 6e shows that NONO and FAPR1 are positively correlated in expression under normal tissues, while they are negatively correlated under tumor tissues, which is reverse against the second and fourth ranked gene pairs. However, these pathways are totally different from the second and fourth pathways, except TP53 $\rightarrow$ CD44, and so these pathways might explain the switching mechanism of FARP1 and NONO. Table 5 shows the biweight midcorrelation between CD44 and each of the two genes in this pair. From the table, we can see that under normal tissues CD44 is positively correlated with both NONO and FAPR1, while under tumor tissues CD44 is positively and negatively correlated with NONO and FAPR1, respectively. This result is consistent with our finding that NONO and FAPR are cooperatively expressed under normal tissues while they are alternatively expressed under tumor tissues.

Table 5. Biweight midcorrelations in gene expression of the fifth ranked pair

Gene 1	Gene 2	r_1(•: Normal)	r_2 (+: Tumor)
NONO	CD44	0.43	0.45
FAPR1	CD44	0.42	−0.58
TP53	CD44	0.15	−0.49
FAPR1	EZR	0.61	0.49

We then further checked the biweight midcorrelation of two neighboring genes which are in the pathway and not shown by 'binding' and related labels. We then checked TP53 → CD44 (inhibitor) and FAPR1 → EZR (domain). The results are shown in Table 5, in which TP53 and CD44 are negatively correlated under tumor tissues and EZR and FAPR1 are positively correlated under both two conditions. This result is also consistent with the labels assigned to these neighboring genes.

Overall, to keep a switching mechanism with parallel pathways, two genes in each pair should have the following correlation with the destination gene. Two paired genes are both positively or negatively correlated with the destination gene under one condition, while under the other condition, two paired genes are correlated with the destination gene in two different ways, i.e. one being positive and the other being negative. In fact, all four pairs, i.e. the top, second, fourth and fifth ranked gene pairs have such correlations. This result also implies that the switching mechanisms found by ROS-DET are reliable.

The detail of the sixth to 10th ranked gene pairs is shown in the Supplementary Data due to space limitations. We found that each of the seventh to 10th gene pairs has parallel pathways, meaning that totally eight out of the top 10 gene pairs had parallel pathways.

DISCUSSION AND CONCLUSION

We have developed an efficient and robust method, ROS-DET, for detecting switching mechanisms in gene expression. ROS-DET clearly outperformed current approaches in a variety of experimental settings. Particularly under the case of expression values with a very small range, where the performance of all competing methods was almost equal to random guessing, ROS-DET achieved a significantly better accuracy. We examined the literature on the top five pairs ranked by ROS-DET and found that each pair has been involved with a biological pathway, which can connect two genes of the pair. Furthermore, four out of the top five pairs have parallel pathways, which were suggested by Li (5) as a typical case of switching mechanisms, implying that four pairs have real switching mechanisms. A possible explanation on the parallel pathways which come to (or start with) the destination gene is that the destination gene controls (or is controlled by) two cases: two pathways are cooperatively (or positively) correlated or two pathways are alternatively (or negatively) correlated. In fact, in each of all four pairs, the biweight midcorrelation between the destination gene and two genes in the pair was consistent with the above explanation. This result also supports the performance of ROS-DET in detecting switching mechanisms in gene expression.

Although the real computation time is not shown in our experimental results, ROS-DET is very time efficient, because the time complexity of computing the biweight midcorrelation coefficient is totally the same as that of a simple correlation coefficient, such as the Pearson's correlation coefficient.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Funding for open access charge: Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Agency (JST).

Conflict of interest statement. None declared.

REFERENCES

1. Pan,W. (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, **18**, 546.
2. Stuart,J.M., Segal,E., Koller,D. and Kim,S.K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249.
3. Ho,Y., Cope,L., Dettling,M. and Parmigiani,G. (2007) Statistical methods for identifying differentially expressed gene combinations. *Methods Mol. Biol.*, **408**, 171.
4. Cho,S.B., Kim,J. and Kim,J.H. (2009) Identifying set-wise differential co-expression in gene expression microarray data. *BMC Bioinformatics*, **10**, 109.
5. Li,K.C. (2002) Genome-wide coexpression dynamics: theory and application. *Proc. Natl Acad. Sci. USA*, **99**, 16875.
6. Dettling,M., Gabrielson,E. and Parmigiani,G. (2005) Searching for differentially expressed gene combinations. *Genome Biol.*, **6**, R88.
7. Watson,M. (2006) CoXpress: differential co-expression in gene expression data. *BMC Bioinformatics*, **7**, 509.
8. Kayano,M., Takigawa,I., Shiga,M., Tsuda,K. and Mamitsuka,H. (2009) Efficiently finding genome-wide three-way gene interactions from transcript- and genotype-data. *Bioinformatics*, **25**, 2735.
9. Ayer,D.E. and Eisenman,R.N. (1993) A switch from Myc: Max to Mad: Max heterocomplexes accompanies monocyte/macrophage differentiation. *Genes Dev.*, **7**, 2110.
10. Lazar,M.A. (2003) Thyroid hormone action: a binding contract. *J. Clin. Invest.*, **112**, 497–499.
11. Cordell,H.J. (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.*, **11**, 2463.
12. Wilcox,R.R. (2005) *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press, San Diego.
13. Shedden,K. and Taylor,J. (2004) Differential correlation detects complex associations between gene expression and clinical outcomes in lung adenocarcinomas. *Methods Microarray Data Anal.*, **4**, 121–131.
14. Hardin,J., Mitani,A., Hicks,L. and VanKoten,B. (2007) A robust measure of correlation between two genes on a microarray. *BMC Bioinformatics*, **8**, 220.
15. Lax,D.A. (1985) Robust estimators of scale: finite-sample performance in long-tailed symmetric distributions. *J. Am. Stat. Assoc.*, **80**, 736–741.

16. Hoaglin,D.C., Mosteller,F. and Tukey,J.W. (1983) *Understanding Robust and Exploratory Data Analysis*. Wiley, New York.
17. Paul,S.R. (1989) Test for the equality of several correlation coefficients. *Can. J. Stat.*, **17**, 217–227.
18. Pearson,K. (1933) On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika*, **25**, 379.
19. Wilcox,R.R. and Muska,J. (2002) Comparing correlation coefficients. *Commun. Stat. Simul. Comput.*, **31**, 49–59.
20. Bartlett,M.S. (1937) Properties of sufficiency and statistical tests. *Proc. R. Soc. Lond. A.*, **160**, 268–282.
21. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M. and Edgar,R. (2007) NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res.*, **35**, D760.
22. Bavner,A., Matthews,J., Sanyal,S., Gustafsson,J.A. and Treuter,E. (2005) EID3 is a novel EID family member and an inhibitor of CBP-dependent co-activation. *Nucleic Acids Res.*, **33**, 3561.
23. Toledo,F. and Wahl,G.M. (2006) Regulating the p53 pathway: in vitro hypotheses, in vivo veritas. *Nat. Rev. Cancer*, **6**, 909–923.
24. Li,M., Brooks,C.L., Kon,N. and Gu,W. (2004) A dynamic role of HAUSP in the p53-Mdm2 pathway. *Mol. Cell*, **13**, 879–886.
25. Li,M., Chen,D., Shiloh,A., Luo,J., Nikolaev,A.Y., Qin,J. and Gu,W. (2002) Deubiquitination of p53 by HAUSP is an important pathway for p53 stabilization. *Nature*, **416**, 648–653.
26. Peterson,D., Lee,J., Lei,X.C., Forrest,W.F., Davis,D.P., Jackson,P.K. and Belmont,L.D. (2010) A chemosensitization screen identifies TP53RK, a kinase that restrains apoptosis after mitotic stress. *Cancer Res.*, **70**, 6325–6335.
27. Dube,D.H. and Bertozzi,C.R. (2005) Glycans in cancer and inflammation potential for therapeutics and diagnostics. *Nat. Rev. Drug Discov.*, **4**, 477–488.
28. Fuster,M.M. and Esko,J.D. (2005) The sweet and sour of cancer: glycans as novel therapeutic targets. *Nat. Rev. Cancer*, **5**, 526–542.
29. Sasisekharan,R., Shriver,Z., Venkataraman,G. and Narayanasami,U. (2002) Roles of heparan-sulphate glycosaminoglycans in cancer. *Nat. Rev. Cancer*, **2**, 521–528.
30. Blackhall,F.H., Merry,C.L.R., Davies,E.J. and Jayson,G.C. (2001) Heparan sulfate proteoglycans and cancer. *Br. J. Cancer*, **85**, 1094.
31. Atsumi,S., Nosaka,C., Ochi,Y., Iinuma,H. and Umezawa,K. (1993) Inhibition of experimental metastasis by an {alpha} -glucosidase inhibitor, 1, 6-epi-cyclophellitol. *Cancer Res.*, **53**, 4896.
32. Pili,R., Chang,J., Partis,R.A., Mueller,R.A., Chrest,F.J. and Passaniti,A. (1995) The {alpha}-glucosidase I inhibitor castanospermine alters endothelial cell glycosylation, prevents angiogenesis, and inhibits tumor growth. *Cancer Res.*, **55**, 2920.
33. Baldwin,T.A. and Ostergaard,H.L. (2001) Developmentally regulated changes in glucosidase II association with, and carbohydrate content of, the protein tyrosine phosphatase CD45. *J. Immunol.*, **167**, 3829.
34. Salmond,R.J., Filby,A., Qureshi,I., Caserta,S. and Zamoyska,R. (2009) T-cell receptor proximal signaling via the Src-family kinases, Lck and Fyn, influences T-cell activation, differentiation, and tolerance. *Immunol. Rev.*, **228**, 9–22, 2009.
35. Taher,T.E.I., Smit,L., Griffioen,A.W., Schilder-Tol,E.J.M., Borst,J. and Pals,S.T. (1996) Signaling through cd44 is mediated by tyrosine kinases. association with p56lck in t lymphocytes. *J. Biol. Chem.*, **271**, 2863–2867.
36. Hernaiz,M., Liu,J., Rosenberg,R.D. and Linhardt,R.J. (2000) Enzymatic modification of heparan sulfate on a biochip promotes its interaction with antithrombin III. *Biochem. Biophys. Res. Commun.*, **276**, 292–297.
37. Liu,J. and Pedersen,L.C. (2007) Anticoagulant heparan sulfate: structural specificity and biosynthesis. *Appl. Microbiol. Biotechnol.*, **74**, 263–272.
38. Pellegrini,L., Burke,D.F., Von Delft,F., Mulloy,B. and Blundell,T.L. (2000) Crystal structure of fibroblast growth factor receptor ectodomain bound to ligand and heparin. *Nature*, **407**, 1029–1034.
39. Mizukoshi,E., Suzuki,M., Loupatov,A., Uruno,T., Hayashi,H., Misono,T., Kaul,S.C., Wadhwa,R. and Imamura,T. (1999) Fibroblast growth factor-1 interacts with the glucose-regulated protein GRP75/mortalin. *Biochem. J.*, **343(Pt 2)**, 461.
40. Wadhwa,R., Yaguchi,T., Hasan,M.K., Mitsui,Y., Reddel,R.R. and Kaul,S.C. (2002) Hsp70 family member, mot-2/mthsp70/ GRP75, binds to the cytoplasmic sequestration domain of the p53 protein. *Exp. Cell Res.*, **274**, 246–253.
41. Godar,S., Ince,T.A., Bell,G.W., Feldser,D., Donaher,J.L., Bergh,J., Liu,A., Miu,K., Watnick,R.S., Reinhardt,F. *et al.* (2008) Growth-inhibitory and tumor-suppressive functions of p53 depend on its repression of CD44 expression. *Cell*, **134**, 62–73, 2008.
42. Xu,X.N. and Screaton,G. (2001) HIV-1 Nef: negative effector of Fas? *Nat. Immunol.*, **2**, 384–386.
43. Seto,E., Usheva,A., Zambetti,G.P., Momand,J., Horikoshi,N., Weinmann,R., Levine,A.J. and Shenk,T. (1992) Wild-type p53 binds to the TATA-binding protein and represses transcription. *Proc. Natl Acad. Sci. USA*, **89**, 12028.
44. Giangrande,P.H., Hallstrom,T.C., Tunyaplin,C., Calame,K. and Nevins,J.R. (2003) Identification of E-box factor TFE3 as a functional partner for the E2F3 transcription factor. *Mol. Cell. Biol.*, **23**, 3707.
45. Rossow,K.L. and Janknecht,R. (2003) Synergism between p68 RNA helicase and the transcriptional coactivators CBP and p300. *Oncogene*, **22**, 151–156.
46. An,W., Kim,J. and Roeder,R.G. (2004) Ordered cooperative functions of PRMT1, p300, and CARM1 in transcriptional activation by p53. *Cell*, **117**, 735–748.
47. Martin,T.A., Harrison,G., Mansel,R.E. and Jiang,W.G. (2003) The role of the CD44/ezrin complex in cancer metastasis. *Crit. Rev. Oncol. Hematol.*, **46**, 165–186, 2003.